

Foundations of AI-Driven Cybersecurity

MSCS640: Artificial Intelligence and Cyber Security

Dr. Mouatez Karbab



أكاديمية جوعان بن جاسم للدراسات الدفاعية
Joaan Bin Jassim Academy for Defence Studies

September 7, 2026



Outline

- 1 From Old Rules to New Intelligence
- 2 Core Security Concepts
- 3 Understanding the Adversary
- 4 Data: The Fuel for AI
- 5 The Modern SOC



Learning Objectives

By the end of this chapter, you will be able to:

- **Explain** the limitations of signature-based security using the "Pyramid of Pain."
- **Compare** Supervised and Unsupervised learning in a security context.
- **Apply** the Cyber Kill Chain framework to analyze real-world breaches.
- **Assess** data quality challenges and the necessity of feature extraction.
- **Evaluate** AI model performance using the Confusion Matrix and error types.



The Shift in Cybersecurity

The Old Paradigm

Security systems were rigid, like a guard checking a list. If a threat wasn't on the list (signature), it was allowed in.

The Modern Reality

Attackers are automated, fast, and sophisticated. They steal credentials and exploit "open windows." Traditional, rule-based defenses are no longer sufficient.

- We must move to **AI-powered defense**.
- **Requirement:** An AI model is only as good as its data. Understanding the "how" of cyber attacks is prerequisite to building the "what" of AI tools.

Outline

- 1 From Old Rules to New Intelligence
- 2 Core Security Concepts
- 3 Understanding the Adversary
- 4 Data: The Fuel for AI
- 5 The Modern SOC



The Limit of Signatures

- **Signature-Based Detection:** Relies on identifying unique patterns (Hashes like MD5/SHA-256).
- **Process:** Identify virus → Calculate Hash → Add to Database → Scan & Block.

The Problem: Polymorphic Malware

Attackers use tools to slightly alter code every time it spreads. This changes the Hash. To a signature-based system, it looks like a completely new, safe file.



The AI Advantage: Behavioral Analysis

AI and Machine Learning shift [2] the focus from "Bad Signatures" to "Bad Behavior."

Concept: The Baseline

The AI observes the network to calculate "normal" activity.

- **Normal:** User A logs in at 9:00 AM from London.
- **Anomaly:** User A logs in at 3:00 AM from a different country.
- **Result:** Detection of **Zero-Day Attacks** (exploits never seen before) based on deviation from the baseline [3].

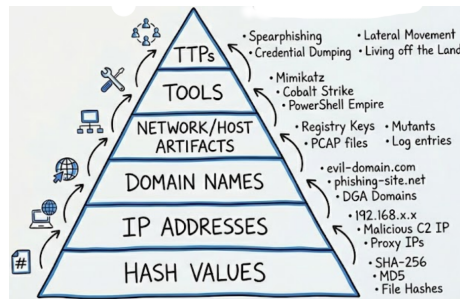
The Pyramid of Pain I

Created by David Bianco, this concept ranks indicators by how much "pain" we cause the attacker if we block them.

- **Bottom (Hash Values & IPs):** Easy for attackers to change. Blocking these causes **No Pain**.
- **Top (TTPs - Behavior):** Tactics, Techniques, and Procedures. This is *how* the attacker operates.

AI's Position

AI operates at the **Top of the Pyramid**. It is very difficult for attackers to unlearn their methods or skills. Detecting behavior causes **Maximum Pain**.



Supervised vs. Unsupervised Learning

Supervised Learning (By Example)

- **Concept:** Teaching with labels ("This is malware", "This is safe").
- **Best For:** Malware detection and classifying known attacks.

Unsupervised Learning (By Observation)

- **Concept:** Feeding raw data without labels. The AI groups similarities and highlights outliers.
- **Best For:** Insider threats and hunting for new, unknown attacks.



Section 1: Assessment (1/3)

Question 1

Why is signature-based detection often ineffective against modern threats?

- A. It consumes too much processing power.
- B. It cannot detect Polymorphic Malware which changes its hash.
- C. It only works on Windows operating systems.
- D. It requires an active internet connection to function.



Section 1: Assessment (1/3) - Answer

Question 1

Why is signature-based detection often ineffective against modern threats?

- A. It consumes too much processing power.
- B. It cannot detect Polymorphic Malware which changes its hash. (✓)
- C. It only works on Windows operating systems.
- D. It requires an active internet connection to function.

Explanation:

Signature-based detection relies on static fingerprints (hashes). Polymorphic malware changes its code structure (and thus its hash) with every iteration, rendering the static signature useless.



Section 1: Assessment (2/3)

Question 2

According to the "Pyramid of Pain," which indicator causes the most pain to an attacker when detected and blocked?

- A. IP Addresses.
- B. Hash Values.
- C. Domain Names.
- D. TTPs (Tactics, Techniques, and Procedures).



Section 1: Assessment (2/3) - Answer

Question 2

According to the "Pyramid of Pain," which indicator causes the most pain to an attacker when detected and blocked?

- A. IP Addresses.
- B. Hash Values.
- C. Domain Names.
- D. TTPs (Tactics, Techniques, and Procedures). (✓)

Explanation:

TTPs represent the attacker's behaviors and skills. While an attacker can easily change an IP address or a file hash, changing their fundamental methods and tools requires significant time and retraining, causing "Maximum Pain."



Section 1: Assessment (3/3)

Question 3

Which type of AI learning is best suited for detecting insider threats where the specific attack pattern is unknown?

- A. Supervised Learning.
- B. Unsupervised Learning.
- C. Reinforcement Learning.
- D. Signature Matching.



Section 1: Assessment (3/3) - Answer

Question 3

Which type of AI learning is best suited for detecting insider threats where the specific attack pattern is unknown?

- A. Supervised Learning.
- B. Unsupervised Learning. (✓)
- C. Reinforcement Learning.
- D. Signature Matching.

Explanation:

Unsupervised learning learns by observation without labels. It establishes a baseline of normal behavior and identifies statistical outliers (anomalies), making it ideal for finding hidden or unknown threats like a rogue insider.



Outline

- 1 From Old Rules to New Intelligence
- 2 Core Security Concepts
- 3 Understanding the Adversary
- 4 Data: The Fuel for AI
- 5 The Modern SOC



The CIA Triad

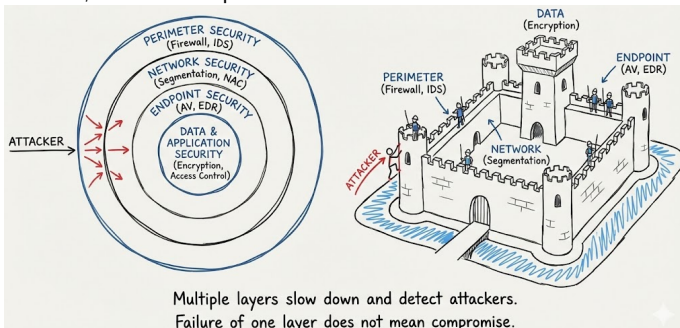
Every security control and AI model protects one of these three goals:

- 1 **Confidentiality:** Secrecy.
 - *AI Defense:* Detecting data exfiltration.
- 2 **Integrity:** Accuracy and Trustworthiness.
 - *AI Defense:* Detecting unauthorized file modifications.
- 3 **Availability:** Systems work when needed.
 - *AI Defense:* Predicting failures or filtering DDoS traffic.

Defense in Depth I

The "Layered Approach" ensures that if one control fails, another is in place.

- **Perimeter:** Firewalls (Network traffic).
- **Endpoint:** Antivirus (Laptops/Servers).
- **Data:** Encryption (Files).



AI's Role

AI is not a replacement for these layers. It is an intelligence layer that *monitors* all other layers simultaneously to connect the dots.

The Attack Surface

The sum of all points where an attacker can try to enter.

- **Network Surface:** Open ports, Wi-Fi, Cloud configurations.
- **Human Surface:** Often the weakest link. Susceptible to **Social Engineering** (Phishing).
- **Software Surface:** Vulnerabilities (bugs) in installed applications.



Section 2: Assessment (1/3)

Question 1

In the CIA Triad, which concept ensures that data has not been altered or deleted by unauthorized individuals?

- A. Confidentiality.
- B. Integrity.
- C. Availability.
- D. Authorization.



Section 2: Assessment (1/3) - Answer

Question 1

In the CIA Triad, which concept ensures that data has not been altered or deleted by unauthorized individuals?

- A. Confidentiality.
- B. Integrity. (✓)
- C. Availability.
- D. Authorization.

Explanation:

Integrity is the guarantee that information is accurate and trustworthy. If an attacker modifies a bank transaction or deletes a log file, they have violated the system's Integrity.



Section 2: Assessment (2/3)

Question 2

What is the core principle of "Defense in Depth"?

- A. Relying solely on AI to catch all threats.
- B. Using a single, very strong firewall.
- C. Using multiple layers of security controls.
- D. Disconnecting the network from the internet.



Section 2: Assessment (2/3) - Answer

Question 2

What is the core principle of "Defense in Depth"?

- A. Relying solely on AI to catch all threats.
- B. Using a single, very strong firewall.
- C. Using multiple layers of security controls. (✓)
- D. Disconnecting the network from the internet.

Explanation:

Defense in Depth uses a layered approach (like a castle with a moat, walls, and guards). If an attacker bypasses the perimeter (firewall), the endpoint protection or data encryption serves as a backup defense.



Section 2: Assessment (3/3)

Question 3

Phishing emails that trick employees into revealing passwords target which part of the attack surface?

- A. Network Surface.
- B. Software Surface.
- C. Human Surface.
- D. Physical Surface.



Section 2: Assessment (3/3) - Answer

Question 3

Phishing emails that trick employees into revealing passwords target which part of the attack surface?

- A. Network Surface.
- B. Software Surface.
- C. Human Surface. (✓)
- D. Physical Surface.

Explanation:

The Human Surface refers to the employees and users within an organization. Social engineering attacks like phishing exploit human psychology (trust, fear, urgency) rather than technical software bugs.



Outline

- 1 From Old Rules to New Intelligence
- 2 Core Security Concepts
- 3 Understanding the Adversary
- 4 Data: The Fuel for AI
- 5 The Modern SOC



The Cyber Kill Chain (Lockheed Martin)

A linear, step-by-step model of a cyber attack. Breaking any link stops the attack.

- 1 **Reconnaissance:** Researching the target.
- 2 **Weaponization:** Creating the malicious payload.
- 3 **Delivery:** Sending the weapon (e.g., email).
- 4 **Exploitation:** Triggering the vulnerability.
- 5 **Installation:** Installing a backdoor.
- 6 **Command and Control (C2):** Remote control of the victim.
- 7 **Actions on Objectives:** Data theft or destruction.



Case Study: Target Corporation (2013)

The Breach Path

- **Recon:** Attackers identified an HVAC vendor (Fazio Mechanical).
- **Delivery/Exploit:** Phished the vendor, stealing credentials.
- **Lateral Movement:** Used vendor credentials to access Target's portal, then jumped to the main network.
- **Objective:** Installed malware on POS registers to steal 40 million credit cards.

Lesson

Attackers rarely attack the front door directly. AI is needed to detect unusual lateral movement (e.g., Vendor Portal → Main Network).



MITRE ATT&CK Framework I

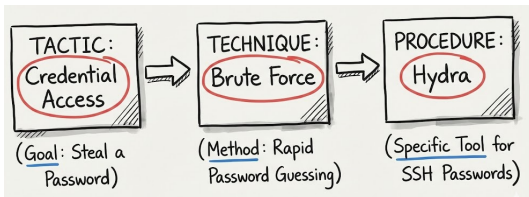
A "Periodic Table" for hacker behavior. It provides the **labels** for AI training.

Reconnaissance	Resource Development	Initial Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Lateral Movement	Collection	Command and Control
11 techniques	8 techniques	11 techniques	17 techniques	23 techniques	14 techniques	47 techniques	17 techniques	34 techniques	9 techniques	17 techniques	18 techniques
Active Scanning (3)	Acquire Access	Content Injection	Cloud Administration Command	Account Manipulation (7)	Abuse Elevation Control Mechanism (6)	Abuse Elevation Control Mechanism (6)	Adversary-in-the-Middle (4)	Account Discovery (4)	Exploitation of Remote Services	Adversary-in-the-Middle (4)	Application Layer Protocol (5)
Gather Victim Host Information (4)	Acquire Infrastructure (8)	Drive-by Compromise	Command and Scripting Interpreter (13)	BITS Jobs	Access Token Manipulation (5)	Access Token Manipulation (5)	Brute Force (4)	Application Window Discovery	Internal Spearphishing	Archive Collected Data (3)	Communication Through Removable Media
Gather Victim Identity Information (3)	Compromise Accounts (3)	Exploit Public-Facing Application	Container Administration Command	Boot or Logon Autostart Execution (14)	Account Manipulation (7)	BITS Jobs	Credentials from Password Stores (6)	Browser Information Discovery	Lateral Tool Transfer	Audio Capture	Content Injection
Gather Victim Network Information (6)	Compromise Infrastructure (8)	External Remote Services	Deploy Container	Boot or Logon Initialization Scripts (5)	Boot or Logon Autostart Execution (14)	Build Image on Host	Exploitation for Credential Access	Cloud Infrastructure Discovery	Remote Service Session Hijacking (2)	Automated Collection	Cloud Obfuscation
Gather Victim Org Information (4)	Develop Capabilities (4)	Hardware Additions	ESXi Administration Command	Cloud Application Integration	Boot or Logon Initialization Scripts (5)	Debugger Evasion	Deobfuscate/Decode Files or Information	Cloud Service Dashboard	Remote Services (8)	Browser Session Hijacking	Data Encoding (2)
Phishing for Information (4)	Establish Accounts (3)	Phishing (4)	Exploitation for Client Execution	Compromise Host Software Binary	Create or Modify System Process (5)	Delay Execution	Deploy Container	Cloud Service Discovery	Replication Through Removable Media	Clipboard Data	Data Obfuscation (3)
Search Closed Sources (2)	Obtain Capabilities (7)	Replication Through Removable Media	Input Injection	Create Account (3)	Domain or Tenant Policy Modification (2)	Direct Volume Access	Domain or Tenant Policy Modification (2)	Cloud Storage Object Discovery	Software Deployment Tools	Data from Cloud Storage	Dynamic Resolution (3)
Search Open Technical Databases (5)	Stage Capabilities (6)	Supply Chain Compromise (3)	Inter-Process Communication (3)	Create or Modify System Process (5)	Escape to Host	Email Spoofing	Email Spoofing	Container and Resource Discovery	Debugger Evasion	Data from Configuration Repository (2)	Encrypted Channel (2)
Search Open Websites/Domains (3)		Trusted Relationship	Native API	Event Triggered Execution (18)	Event Triggered Execution (18)	Execution Guardrails (2)	Execution Guardrails (2)	Device Driver Discovery	Device Driver Discovery	Data from Information Repositories (6)	Fallback Channels
Search Threat Vendor Data		Valid Accounts (4)	Poisoned Pipeline Execution	Exclusive Control	Exploitation for Privilege Escalation	Exploitation for Defense Evasion	Exploitation for Defense Evasion	Domain Trust Discovery	Use Alternate Authentication Material (4)	Data from Local System	Hide Infrastructure
Search Victim-Owned Websites		Wi-Fi Networks	Scheduled Task/Job (5)	External Remote Services	Hijack Execution Flow (12)	File and Directory Permissions Modification (2)	File and Directory Permissions Modification (2)	File and Directory Discovery	Group Policy Discovery	Data from Network Shared Drive	Ingress Tool Transfer
			Serverless Execution	Hijack Execution Flow (12)	Process Injection (12)	Hide Artifacts (14)	Hide Artifacts (14)	Local Storage Discovery	Log Enumeration	Data from Removable Media	Multi-Stage Channels
			Shared Modules			Hijack Execution Flow (12)	Hijack Execution Flow (12)	Network Sniffing	Network Service	Data Staged (2)	Non-Application Layer Protocol
			Software Deployment Tools			Impair Defenses (12)	Impair Defenses (12)	OS Credential Dumping (8)		Email	Non-Standard Port
			System	Implant Internal	Scheduled						



MITRE ATT&CK Framework II

- **Tactics (The Goal):** *Why* are they doing this? (e.g., Credential Access).
- **Techniques (The Method):** *How* are they doing it? (e.g., Brute Force).
- **Procedures (The Specifics):** *What* tool did they use? (e.g., Hydra).



Importance for AI

It allows us to classify events precisely. Instead of just "Alert," the AI can say "Technique: Brute Force detected."

Section 3: Assessment (1/3)

Question 1

In the Cyber Kill Chain, what is the "Delivery" phase?

- A. Creating the virus.
- B. Transmitting the weaponized payload to the victim (e.g., via email).
- C. Researching the target's employees.
- D. Stealing the data from the database.



Section 3: Assessment (1/3) - Answer

Question 1

In the Cyber Kill Chain, what is the "Delivery" phase?

- A. Creating the virus.
- B. Transmitting the weaponized payload to the victim (e.g., via email). (✓)
- C. Researching the target's employees.
- D. Stealing the data from the database.

Explanation:

Delivery is the transmission phase. It is how the attacker gets the malicious object (created in the Weaponization phase) onto the target's system, commonly through phishing emails, USB drives, or malicious websites.



Section 3: Assessment (2/3)

Question 2

What does the "Tactic" represent in the MITRE ATT&CK framework?

- A. The specific software tool used.
- B. The adversary's technical goal (The "Why").
- C. The method used to achieve the goal (The "How").
- D. The IP address of the attacker.

Section 3: Assessment (2/3) - Answer

Question 2

What does the "Tactic" represent in the MITRE ATT&CK framework?

- A. The specific software tool used.
- B. The adversary's technical goal (The "Why"). (✓)
- C. The method used to achieve the goal (The "How").
- D. The IP address of the attacker.

Explanation:

In MITRE ATT&CK, Tactics represent the high-level objective of the attacker, such as "Initial Access," "Execution," or "Exfiltration." It answers *why* the attacker is performing an action.



Section 3: Assessment (3/3)

Question 3

What was the key "lateral movement" step in the Target Corporation breach?

- A. Moving from the HVAC vendor portal to the main network.
- B. Breaking into the physical store at night.
- C. Moving data from the US to Russia.
- D. Switching from Windows to Linux servers.



Section 3: Assessment (3/3) - Answer

Question 3

What was the key "lateral movement" step in the Target Corporation breach?

- A. Moving from the HVAC vendor portal to the main network. (✓)
- B. Breaking into the physical store at night.
- C. Moving data from the US to Russia.
- D. Switching from Windows to Linux servers.

Explanation:

The attackers compromised a less secure third-party vendor (HVAC) first. They then used those credentials to move laterally (sideways) through the network to reach the high-value target: the Point-of-Sale network.



Outline

- 1 From Old Rules to New Intelligence
- 2 Core Security Concepts
- 3 Understanding the Adversary
- 4 Data: The Fuel for AI
- 5 The Modern SOC



Key Data Sources

- **Network Flows:** The "Phone Bill." Summaries of traffic (Who, When, How much). Fast and lightweight.
- **Packet Captures (PCAP):** The "Full Recording." Every detail. Massive file sizes.
- **Endpoint Logs:** The "Device Diary." Logins, process creation (e.g., Sysmon).
- **Web Logs:** The "Visitor Log." Requests to web servers.



The "Big Data" Challenge

- **Volume:** Terabytes of data per day.
- **Velocity:** Attacks happen in milliseconds; real-time processing is required.
- **Variety:** Different formats (Text logs, JSON, Raw packets).
- **Imbalance:** The biggest hurdle for AI.
 - 999,990 "Safe" events vs. 10 "Attack" events.
 - AI struggles to learn what an attack looks like because examples are rare.



Data Preparation Pipeline I

"Garbage In, Garbage Out." Data must be cleaned before training.

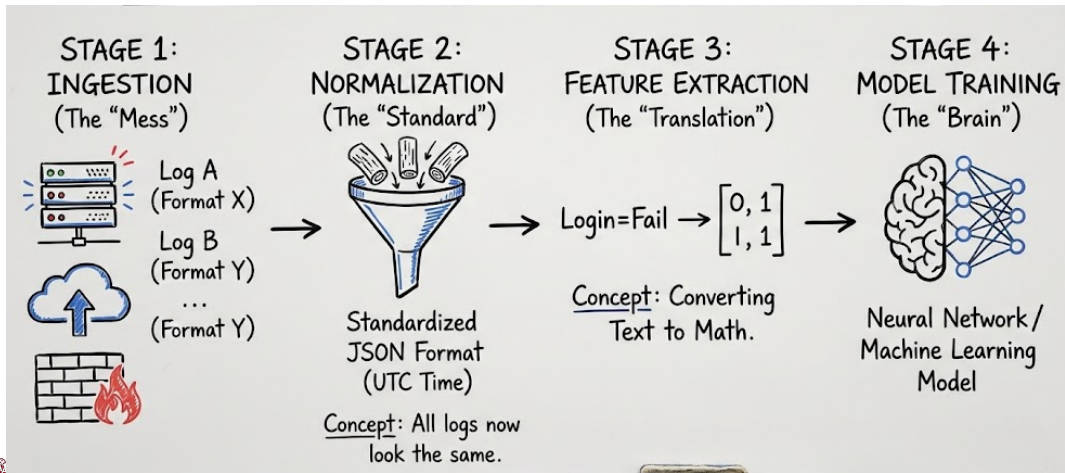
- 1 **Normalization:** Converting different tool languages (Microsoft vs. Cisco) into one standard format (e.g., JSON).
 - **Tool A says:** "User: Admin logged in."
 - **Tool B says:** "Account: Admin, Action: Authentication Success."
- 2 **Time Synchronization:** All devices must use **UTC**. If clocks differ, the cause-and-effect timeline of an attack is destroyed.

Feature Extraction

AI models (Neural Networks) cannot read text. They understand math.

- Raw: "Status=Fail" → Feature: `is_failure` = 1
- Raw: "10:00 AM" → Feature: `hour` = 10

Data Preparation Pipeline II



Python: Data Simulation (1/2)

This code block simulates loading raw log data. In a real scenario, this would be a large CSV file.

```

import pandas as pd
import io

# 1. Simulate raw log data (CSV format)
log_data = """timestamp,source_ip,status_code
2025-08-16 10:00:01,192.168.1.10,200
2025-08-16 10:00:05,192.168.1.10,404
2025-08-16 10:00:10,192.168.1.10,404
2025-08-16 10:00:15,10.0.0.5,200
"""

# 2. Load the data into a DataFrame
df = pd.read_csv(io.StringIO(log_data))
    
```

Python: Feature Extraction Logic (2/2)

Here, we convert raw "404" codes into a mathematical count. An AI can learn that a high count is suspicious.

3. FEATURE EXTRACTION

Logic: Group by IP address, count how many '404' errors appear.

```
features = df.groupby('source_ip')['status_code'].apply(
    lambda x: (x == 404).sum()
).reset_index(name='error_count_feature')
```

```
print("Data ready for AI Model:")
print(features)
```

Output:

```
#      source_ip  error_count_feature
# 0      10.0.0.5                    0
# 1     192.168.1.10                  2
```



Section 4: Assessment (1/3)

Question 1

Which data source provides the "Full Recording" of network traffic, including the content of the conversation?

- A. Network Flows.
- B. Packet Captures (PCAP).
- C. DNS Logs.
- D. Firewall Alerts.



Section 4: Assessment (1/3) - Answer

Question 1

Which data source provides the "Full Recording" of network traffic, including the content of the conversation?

- A. Network Flows.
- B. Packet Captures (PCAP). (✓)
- C. DNS Logs.
- D. Firewall Alerts.

Explanation:

PCAP data captures the full payload and headers of network packets. While Network Flows act like a phone bill (metadata only), PCAP acts like a wiretap recording of the actual conversation.



Section 4: Assessment (2/3)

Question 2

What is "Data Imbalance" in the context of training AI for security?

- A. Having too much data to store.
- B. Having data from too many different time zones.
- C. Having a massive amount of "safe" data and very few examples of "attacks."
- D. Having data that is not encrypted.



Section 4: Assessment (2/3) - Answer

Question 2

What is "Data Imbalance" in the context of training AI for security?

- A. Having too much data to store.
- B. Having data from too many different time zones.
- C. Having a massive amount of "safe" data and very few examples of "attacks." (✓)
- D. Having data that is not encrypted.

Explanation:

Security datasets are naturally skewed. Real attacks are rare compared to normal background traffic. This makes it difficult for AI models to learn the characteristics of an attack effectively without special handling.



Section 4: Assessment (3/3)

Question 3

Why is "Feature Extraction" necessary?

- A. To encrypt the data for safety.
- B. To convert text-based logs into numerical values that Neural Networks can process.
- C. To compress the data to save disk space.
- D. To delete the data after it has been read.



Section 4: Assessment (3/3) - Answer

Question 3

Why is "Feature Extraction" necessary?

- A. To encrypt the data for safety.
- B. To convert text-based logs into numerical values that Neural Networks can process. (✓)
- C. To compress the data to save disk space.
- D. To delete the data after it has been read.

Explanation:

Most machine learning algorithms are mathematical functions that require numerical input. Raw logs contain text (e.g., "Login Failed"). Feature extraction translates this text into numbers (e.g., Failure Count = 1) that the AI can calculate.



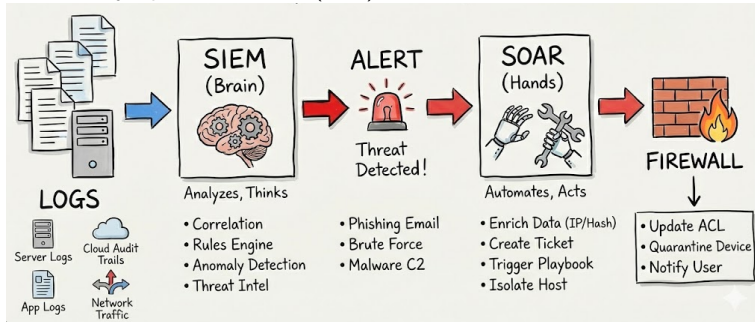
Outline

- 1 From Old Rules to New Intelligence
- 2 Core Security Concepts
- 3 Understanding the Adversary
- 4 Data: The Fuel for AI
- 5 The Modern SOC



The Modern SOC

The Security Operations Center (SOC) combines humans and machines.



- **SIEM (Security Information and Event Management)[1]:** The "Brain." Collects logs and uses AI to analyze them.
- **SOAR (Security Orchestration, Automation, and Response):** The "Hands." Executes actions (e.g., blocking an IP automatically).



The Confusion Matrix

We measure AI success by its errors.

	Actually Malicious	Actually Safe
AI Predicts Threat	True Positive (Success)	False Positive (Noise)
AI Predicts Safe	False Negative (Breach)	True Negative (Quiet)

- **False Positive (Type I):** False Alarm. Leads to analyst "Alert Fatigue."
- **False Negative (Type II):** Missed Attack. The most dangerous outcome.



AI as a Force Multiplier & XAI

Force Multiplier

AI increases output. An analyst can review 500 alerts instead of 50 because AI filters the noise.

The "Black Box" Problem

Complex models often give answers without reasons. If an analyst doesn't understand *why* AI blocked a user, they won't trust it.

- **Explainable AI (XAI):** Ensures models provide reasons (e.g., "Blocked because location changed rapidly").
 - *AI says:* "Block User A."
 - *Analyst asks:* "Why?"
 - *AI:* (Silence).



Section 5: Assessment (1/3)

Question 1

In a Confusion Matrix, what is a "False Negative"?

- A. The AI alerts on a safe file.
- B. The AI correctly identifies a virus.
- C. The AI fails to detect an actual attack (Silence during a breach).
- D. The AI correctly ignores a safe file.



Section 5: Assessment (1/3) - Answer

Question 1

In a Confusion Matrix, what is a "False Negative"?

- A. The AI alerts on a safe file.
- B. The AI correctly identifies a virus.
- C. The AI fails to detect an actual attack (Silence during a breach). (✓)
- D. The AI correctly ignores a safe file.

Explanation:

A False Negative is a Type II error where the system predicts "Safe" but the reality is "Malicious." This is the most dangerous scenario in cybersecurity because an attack proceeds unnoticed.



Section 5: Assessment (2/3)

Question 2

What is the primary function of a SOAR platform in a SOC?

- A. To collect logs.
- B. To automate responses and execute actions (The "Hands").
- C. To train the AI model.
- D. To replace all human analysts.

Section 5: Assessment (2/3) - Answer

Question 2

What is the primary function of a SOAR platform in a SOC?

- A. To collect logs.
- B. To automate responses and execute actions (The "Hands"). (✓)
- C. To train the AI model.
- D. To replace all human analysts.

Explanation:

While the SIEM acts as the "Brain" (analysis), the SOAR acts as the "Hands." It orchestrates tools to take action, such as automatically blocking an IP address on a firewall or disabling a compromised user account.



Section 5: Assessment (3/3)

Question 3

Why is Explainable AI (XAI) critical in a Security Operations Center?

- A. It makes the AI run faster.
- B. It allows the AI to encrypt data better.
- C. It builds trust by telling analysts *why* a decision was made.
- D. It removes the need for data cleaning.



Section 5: Assessment (3/3) - Answer

Question 3

Why is Explainable AI (XAI) critical in a Security Operations Center?

- A. It makes the AI run faster.
- B. It allows the AI to encrypt data better.
- C. It builds trust by telling analysts *why* a decision was made. (✓)
- D. It removes the need for data cleaning.

Explanation:

Analysts suffer from the "Black Box" problem where they cannot verify AI decisions. XAI provides the rationale (e.g., "Blocked due to impossible travel"), allowing humans to trust and verify the automated actions.



Conclusion

- **Evolution:** We have moved from static signatures to dynamic behavioral analysis.
- **Strategy:** We use models like the **Pyramid of Pain** and **Defense in Depth** to structure our defenses.
- **Adversary:** Understanding the **Cyber Kill Chain** and **MITRE ATT&CK** allows us to label data effectively.
- **Data:** AI requires clean, normalized data and **Feature Extraction** to function.
- **Operations:** In the SOC, AI balances False Positives and Negatives, acting as a force multiplier for human analysts.

References I

- [1] Srinivas Reddy Pulyala.
From detection to prediction: Ai-powered siem for proactive threat hunting and risk mitigation.
Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2024.
- [2] Abdullah Al Siam, M. Alazab, A. Awajan, and Nuruzzaman Faruqui.
A comprehensive review of ai's current impact and future prospects in cybersecurity.
IEEE Access, 13:14029–14050, 2025.
- [3] S. Zeadally, Erwin Adi, Zubair A. Baig, and Imran A. Khan.
Harnessing artificial intelligence capabilities to improve cybersecurity.
IEEE Access, 8:23817–23837, 2020.



Acronyms and Terminology I

- AI (Artificial Intelligence)** A field of computer science focused on creating systems that can perform tasks requiring human intelligence. In this context, it refers to systems that can learn normal behavior, detect deviations, and automate defensive actions.
- Anomaly** A deviation from a baseline of established, normal behavior within an IT environment. Anomalies can indicate a potential security threat.
- APT (Advanced Persistent Threat)** A prolonged and highly targeted cyber campaign where an attacker gains unauthorized access to a network and remains undetected for an extended period, typically for espionage or data theft.
- Adversarial examples** Slightly modified malicious inputs (like a malware file with a few bytes changed) specifically designed to be misclassified and evaded by defensive AI systems.
- Cyber Kill Chain** A framework that outlines the typical sequential stages of a cyberattack, suggesting that disrupting the attack at any stage can prevent the adversary's final objective.
- Deepfakes** AI-generated synthetic audio or video used to impersonate trusted individuals, often for sophisticated social engineering schemes.
- Fuzzing** An automated method used by attackers to discover new software vulnerabilities by providing invalid, unexpected, or random data to a program's inputs.



Acronyms and Terminology II

GANs (Generative Adversarial Networks) A type of AI model that can be trained by attackers to automatically create new, functional strains of malware that are invisible to signature-based defenses.

IDS (Intrusion Detection Systems) Security tools that monitor network or system activities for malicious actions or policy violations. Traditional versions often rely on signatures.

IOCs (Indicators of Compromise) Subtle clues, patterns, or forensic data found within logs or files that indicate a security breach has likely occurred.

ML (Machine Learning) A subfield of AI that enables computer systems to learn from data and identify patterns without being explicitly programmed. It is the core technology behind behavior-based threat detection.

MITRE ATT&CK A globally accessible knowledge base and framework of real-world adversary tactics and techniques, used by security professionals to understand and model attacker behavior.

Polymorphic malware Malicious software that constantly changes its code or features to create new, unique signatures each time it replicates, making it difficult for traditional antivirus software to detect.

SIEM (Security Information and Event Management) A central platform that collects, aggregates, and analyzes security-related log data and alerts from across an organization's entire IT infrastructure.



Acronyms and Terminology III

SOAR (Security Orchestration, Automation, and Response) A platform that helps organizations respond to security alerts in an automated fashion by using predefined workflows or "playbooks" to execute response actions.

Signature A unique pattern, like a digital fingerprint, associated with a known piece of malicious code or a specific attack method, used by traditional security tools for detection.

Supply chain attack An indirect attack strategy where adversaries compromise an organization by targeting less secure elements in its network of suppliers, such as a third-party software vendor.

Threat Hunting The proactive practice of searching through an organization's networks and data to find advanced threats that have evaded existing automated security defenses.

Zero-day attack A cyberattack that exploits a previously unknown software vulnerability for which no patch or fix is available, making it highly effective against traditional defenses.

AI Artificial Intelligence: The simulation of human intelligence in machines, programmed to learn, reason, and automate tasks, particularly for analyzing data and making decisions.

API Application Programming Interface: A set of rules and tools that allows different software applications to communicate with each other. Insecure APIs can be a major vulnerability.



Acronyms and Terminology IV

ATT&CK Adversarial Tactics, Techniques, and Common Knowledge: A globally accessible knowledge base and framework from MITRE that categorizes adversary actions based on real-world observations.

Availability A core principle of the CIA Triad ensuring that systems and data are accessible to authorized users when they are needed.

C2 Command and Control: The communication channel an attacker uses to maintain control over a compromised system, issue commands, and extract data.

Checksum A small block of data derived from a larger block, used to detect errors or unauthorized changes to the original data, thus ensuring integrity.

CIA Triad A foundational security model representing the three core goals of information security: Confidentiality, Integrity, and Availability.

Confidentiality A core principle of the CIA Triad ensuring that sensitive information is accessible only to authorized individuals.

Cyber Kill Chain A model that describes the typical sequence of stages an adversary must complete to carry out a cyberattack.

Denial-of-Service (DoS) An attack designed to make a system or service unavailable to legitimate users by overwhelming it with traffic or requests.



Acronyms and Terminology V

- Digital Signature** A cryptographic technique used to verify the authenticity and integrity of a digital message or file, confirming it hasn't been altered.
- EDR** Endpoint Detection and Response: A category of security tools that continuously monitor end-user devices (like laptops and servers) to detect and respond to advanced threats.
- ELF** Executable and Linkable Format: The standard file format for executable programs and libraries on Linux and most other Unix-like operating systems.
- Encryption** The process of converting data into an unreadable format (ciphertext) to protect its confidentiality. A key is required to decrypt it back to readable data.
- False Positive** An alert from a security system that incorrectly identifies normal, benign activity as malicious.
- IAM** Identity and Access Management: The security discipline that ensures the right individuals have the appropriate level of access to technology resources.
- IDS** Intrusion Detection System: A security tool that monitors network or system activities for malicious patterns or policy violations and reports them.
- Integrity** A core principle of the CIA Triad focused on maintaining the accuracy and trustworthiness of data, ensuring it has not been altered without authorization.



Acronyms and Terminology VI

IOCs Indicators of Compromise: Forensic artifacts or pieces of data (like file hashes, IP addresses, or domain names) that indicate a security breach has occurred.

MFA Multi-Factor Authentication: A security mechanism that requires a user to provide two or more different types of verification factors to gain access to an account or system.

MITRE ATT&CK See ATT&CK.

NIST National Institute of Standards and Technology: A U.S. government agency that creates influential standards and frameworks for cybersecurity.

PCAP Packet Capture: A file containing the raw data of packets captured from a network, providing a detailed record of network traffic for analysis.

PE Portable Executable: The standard file format for executable programs and libraries used by the Windows operating system.

SAST Static Application Security Testing: The process of analyzing an application's source code for security vulnerabilities without actually running the code.

Security Control A safeguard or countermeasure, such as a firewall or encryption, put in place to avoid, detect, or minimize security risks.



Acronyms and Terminology VII

- Security Telemetry** A general term for the wide variety of security data (e.g., logs, network flows) collected from across a digital environment for monitoring and analysis.
- SOC** Security Operations Center: A centralized team, along with its associated technology and processes, responsible for continuously monitoring and defending an organization against cyber threats.
- Spoofing** An attack where a person or program successfully impersonates another entity by falsifying data, such as an email sender or an IP address.
- SQL Injection** A web security vulnerability that allows an attacker to interfere with the queries that an application makes to its database, often to extract data.
- STRIDE** A threat modeling framework that categorizes potential threats into six types: Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege.
- Sysmon** A Microsoft tool that provides detailed and high-quality logging of system activity on Windows, such as process creation and network connections.
- Tampering** The unauthorized modification of data, which is a direct threat to data integrity.
- Threat Intelligence** Information and data about current or emerging cyber threats, including details on attackers and their methods, used to inform security decisions.



Acronyms and Terminology VIII

- True Positive** An alert from a security system that correctly identifies a genuine malicious activity or security threat.
- TTPs** Tactics, Techniques, and Procedures: A term used in threat intelligence to describe the behavior and methods of an adversary during an attack.
- XSS** Cross-Site Scripting: A type of web security vulnerability that allows an attacker to inject malicious scripts into web pages viewed by other users.



Thank You!

